Source Oriented Data Processing and Quantification: Distrustful Brothers [1995]

Thaller, Manfred

Historical Social Research
Historische Sozialforschung

Manfred Thaller:

Source Oriented Data Processing and Quantification: Distrustful Brothers [1995]

doi: 10.12759/hsr.suppl.29.2017.287-306

Published in:

Historical Social Research Supplement 29 (2017)

Cite as:

Manfred Thaller. 2017. Source Oriented Data Processing and Quantification: Distrustful Brothers [1995]. *Historical Social Research* Supplement 29: 287-306. doi: 10.12759/hsr.suppl.29.2017. 287-306.

Historical Social Research
Historische Sozialforschung

Other articles published in this Supplement:

Source Oriented Data Processing and Quantification: Distrustful Brothers [1995]

Manfred Thaller *

Abstract: »*Quellenorientierte Datenverarbeitung und Quantifizierung: misstrauische Brüder*«. Historians using computers to apply quantitative methods and historians using computers for other things have strained relationships. This is not the result of mutual bad will, but results from genuine differences between information as provided by contemporary phenomena and that derived from historical documents. Only after these differences are taken care of quantitative and other formal methods are methodologically safe. A tentative model for a possible solution of part of these problems by fuzzy logic is presented. More generally we discuss which developments are needed in the application of information technology to historical research to ease the strained relationship mentioned at the beginning.

Keywords: Fuzzy logic, quantitative methods, epistemology.

1. Introduction

Source oriented data processing and the application of quantitative methods to history are sometimes seen as contradictions, virtually excluding each other. Historically speaking this is very strange, as the roots of source oriented technology are clearly in the tradition of quantitative historical research. In a nutshell, the development roughly ran like this:

When computer applications in historical research started in the fifties and sixties, it was almost inevitable that data were entered in a highly coded way that is in the form of a string of statistical variables, even when the purpose of the project never really went beyond fairly simple counting and sorting operations. This was of course connected with all the classical problems of data entry: avoiding typing errors and checking for the reliability of the coding process. As a result, the seventies saw a number of proposals to enter all variables which were not clearly numerical in the original sources, as strings of characters which would then be converted into appropriate codes for statistical operations by suitable programs. At the begin-

* Reprint of: Manfred Thaller. 1995. Source Oriented Data Processing and Quantification: Distrustful Brothers. In *Statistics for Historians: Standard Packages and Specific Historical Software*, ed. Manfred Thaller et al., 125-44. St. Katharinen: Scripta Mercaturae (= Halbgraue Reihe zur Historischen Fachinformatik A 26).

ning, these were usually embodied in specialized coding programs[1]: at a later stage, it was frequently argued that such spezialized programs would not be needed anymore, as the recoding facilities of the larger statistical packages on the one hand, the appropriate tools of data base packages on the other, would allow one to do so in a general purpose environment just as well.

So, if source oriented data processing is deeply rooted within the practical requirements of quantitative studies, why is it that in the meantime it is almost seen as hostile to quantification, as a step back from the formalizations which the quantitative enthusiasm of the seventies saw as the guiding reason for the application of computers in history?

Well many quantifiers probably misunderstood the reactions of some of their colleagues in the seventies: a huge majority followed the advice, about how to apply computers to historical research *not,* because their quantitative colleagues convinced them, but because at that time a more or less close embrace with statistics was the only way to get the beastly mainframes of the time to do the sorting and searching, which was all a large portion of the historians using "statistics" at that time actually wanted to perform.

And the fact that today the majority of computer-using historians, as opposed to the seventies, are *not* quantifiers, sometimes even emphatically non-quantifiers, is not the result of a devious conspiracy of source oriented methodology, seducing the innocent historical community away from chaste statistics into the sensual embrace of non-quantitative argumentation: that embrace simply is what a huge majority of them wanted from the beginning.

The author of this paper finds himself in a very strange position. Being today considered very much a proponent of source oriented computing, I started my computer related work with applying and teaching quantitative methods, advocating the use of fuzzy set theory in history[2] in the early eighties and applying cluster analysis to the history of mentality at the same time[3].

Looking back at my arguments over the last twenty years, I discover that, during my first years in computing, just as everybody else, I considered source oriented computing very much a tool to prepare for a final, crowning analysis. During the first half of the eighties, I regularly produced papers and statements which argued the point that, by developing specific programming tools within source oriented computing, statistical analysis could be applied to an incomparably larger domain within history than so far. In about 1984 I quite abruptly and almost completely

---

1 E.g.: Donald E. Ginter et al., 'A Review of Optimal Input Methods: Fixed Field, Free Field, and the Edited Text' *Historical Methods Newsletter* 10 (1977), 166-76; Mark Overton 'Computer Analysis of an Inconsistent Data Source: the Case of Probate Inventories' *Journal of Historical Geography* 3 (1977), 317-26; T. J. King 'The Use of Computers for Storing Records in Historical Research' *Historical Methods* 14 (1981), 59-64.

2 Manfred Thaller 'Ungefähre Exaktheit. Theoretische Grundlagen und praktische Möglichkeiten einer Formulierung historischer Quellen als Produkte "unscharfer" Systeme.' *Neue Ansätze in der Geschichtswissenschaft.* Ed. H. Nagl-Docekal and F. Wimmer. Wien: VWGÖ, 1984 (= *Conceptus Studien* 1), 77-100. Reprinted in this HSR Supplement, p. 138-159.

3 Manfred Thaller 'Zur Formalisierbarkeit hermeneutischen Verstehens in der Historie.' *Mentalitäten und Lebensverhältnisse. Beispiele aus der Sozialgeschichte der Neuzeit. Rudolf Vierhaus zum 60. Geburtstag.* Göttingen: Vandenhoeck & Ruprecht 1982, 439-54.

stopped doing so: which indeed I myself discovered only when preparing this paper.

In this, I believe, I have been quite typical of historians subscribing to source oriented computing. In this paper, I would like to explore this development in three steps:

- I would like to re-visit one of the conceptual proposals which I made to a research project on historical social mobility in 1978.
- Considerably shorter, I would like then to summarize why I, and as far as I can see many other people in similar positions, put such considerations into the background of their endeavors.
- Finally I would like to argue that the conceptual developments which were undertaken in these early years, for structural reasons could (and should) now be continued with much more promise than formerly.

2. Two Brothers United: Quantification and Source Oriented Computing, AD 1978, or: On the Applicability of Probabilistic Concepts in Research Dealing with Social Phenomena of "Long" Duration

2.1 Background and Formalisms

Giving a title to a research paper[4] like the one chosen, forces us to start with at least a preliminary definition: Social phenomena are of a "long duration" in our sense, if it is impossible to collect all information which is required to investigate them meaningfully under the immediate control of a single person or team.

This definition remains intentionally vague – what constitutes "immediate control" seems to be difficult to define abstractly, so we will not even try it. Instead of that, we will show which problems occur in cases, where it is missing unmistakably.

Control over the collection of information on social processes is clearly missing in cases where data are analyzed which have been collected for other purposes than their analysis within one of the social sciences. Such analyses of data which are process-produced in the broadest sense possible, will always be necessary when we either investigate mechanisms of societies gone by, or when we try to follow a phenomenon which we observe within the current society into the past which cannot be directly observed anymore.

A typical example for such an approach is the research on processes of social mobility, as it has been undertaken for the 19th century by a historical school which

[4] The following is a slightly modified excerpt from an internal research paper from 1978. The notes pointing to the sociological literature of the time have been deleted, except where the point to the specific examples of historical sociology from which I started. In some cases where similarities between my reasoning and articles published later in the area of fuzzy systems occurred, I have added them in footnotes.

understands itself as *Historische Sozialwissenschaft*[5] (historical social research) as well as by sociology herself[6]. Usually sources from these times are used which resemble lists reporting on the social status of persons involved in regularly re-occurring events – like the marriage registers of the religious communities or the secular administration. Such lists contain usually occupational terms, which are used to estimate the social position of the persons mentioned; if they also contain similar occupational information for the fathers of the persons mentioned, all conditions seem to be fulfilled for it to be possible to make statements about the "mobility" between two specific generations.[7]

If we do so, we make an assumption which is so trivial that it is scarcely made explicit. We assume that the fact that a person x belongs to the occupational group G, would imply that the probability $p^x_{GG'}$, with which the father of *this* person belongs to another occupational group G' is equal to the general probability $p_{GG'}$ with which *any* person from group G has a father from group G'. More formally:

$$x \in G \Rightarrow p^x_{GG'} = p_{GG'} \qquad (1)$$

Or becoming seemingly even more trivial, we make the assumption that a person who has an occupation designed by the occupational *term* γ actually belongs to the occupational group G:

$$\gamma \in G, \forall \gamma \qquad (2)$$

Notationally we will henceforth use lower case Greek to denote linguistic terms, hinting at, for example, a specific occupation, and uppercase Greek to specify the set of all such terms occurring within our source. Γ is therefore the set of all γ – for example of all blacksmiths – occurring in our source. Uppercase Roman letters, on the other hand, do not signify anything found in the sources, but a past reality – as, for example, the abstract social position 'self-employed in the processing of metal'.

Strictly speaking, our first assumption is independent from the second one – we started with the implicit idea that we would know to which occupational group a person belonged. We will discuss further below a problem which still occurs with this first assumption, but quite independent from that it is obvious, that formalism (2) has to be fulfilled for formalism (1) to hold. This second assumption is violated with great regularity by all data sets, however, which we discuss here: a typical example is occupational terms which can indicate a person employed by an artisan

---

[5] At that time one of the central areas of quantification within historical research in Germany. See the relatively large number of projects quoted in: Wolfgang Bick et al. (Eds.) *QUANTUM Dokumentation: Quantitative Historische Forschung 1977,* Stuttgart 1977 (=*Historisch Sozialwissenschaftliche Forschungen* 1), <http://www.ssoar.info/ssoar/handle/document/32547>.

[6] One should point out that Natalie Rogoff (*Recent Trends in Occupational Mobility*, Glencoe, 1953), as well as Renate Mayntz (*Soziale Schichtung und sozialer Wandel in einer Industriegemeinde,* Stuttgart 1958), two of the "classics" of the field, are based upon sources as discussed here. There the applied categories remain so broad, however, that the problems discussed below have scarcely been noticed.

[7] A typical example for the completely uncritical use of such data, which has led to extremely severe criticism by historians, are Kleining's two articles 'Struktur- und Prestigemobilität in der BRD' *KZfSS* 23 (1971), 1-32 and 'Soziale Mobilität in der BRD' *KZfSS* 27 (1975), 97-121, 273-92.

just as well as a person employed in industrially organized enterprises. A property of the systems of occupational terminology in natural language, which has to be dealt with in any research which deals with data from the not absolutely immediate past.

A non-historian might claim that we have introduced an artificial difficulty here, as we have not yet precisely defined an "occupational group" within a meta-language and have therefore been caught by the ambiguities of natural language. We simply have to *define* that employment within an industrial establishment is more central for inclusion within a specific "occupational group" than the fact that another person employed by an artisan has superficially performed the same type of work. It is quite obvious that a "blacksmith" who has been employed in the early iron industry does not belong into the same socially relevant group as the "blacksmith" who worked as journeyman for the local smith's shop, which probably continued to exist for a couple of decades at the same location as such an early industrial establishment.

Unfortunately in data we have *not* collected ourselves, the source simply says "blacksmith" in both cases, not "smith working in an industrial establishment" in the one and "blacksmith journeyman" in the other.

A tentative solution of the described problem might be the following: we collect information on how many of the blacksmiths our potential iron works employed, how many smith's shops there were in the town in question and how many blacksmith journeymen one of these typically employed. If we have the proportions between these numbers of employed persons, we can assume with reason that a specific person x with an occupational term γ did belong to an occupational group G with a specific probability: that probability being described by the proportion of all persons out of the set of persons with the occupational term γ which belonged to the specified group G. Which we would like to formalize in the following way:

$$x = \gamma \Rightarrow p_G^x = \frac{n_{\{\gamma|\gamma \in G\}}}{n_T} \qquad (3)$$

Two explanations to this form of notation, which we will use throughout. The general expressions of this form we would like to read as: "The fact that x has the property γ implies a specific probability for x being a member of G, which is given by the third expression."

The superscripted n in front of the sets will describe the number of elements within this set. We introduce this notational form here, because later on we will want to express the fact that sets may have "qualities"[8], which go beyond the characteristics of the single elements which are used to define the set in the first place, but which are equally derived from properties of the elements constituting it. These "qualities" will henceforth always be expressed by such preceding superscripts, with the most general form of $*\{m_1, \ldots, m_i\}$ or $*M$ representing the set of qualities that can be defined for the elements of the set M.

[8] See in the meantime the proposition of Anio O. Arigoni in 'Mathematical Developments Arising from "Semantic Implication" and the Evaluation of Membership Characteristic Functions', *Fuzzy Sets and Systems* 4 (1980), 167-83, particularly the reasoning on page 167 from which he starts.

Starting with formalism (3) we can of course use that notion of the "quality" of a set to express our last statement more simply and at the same time more generally. We say now, that the probability with which the fact that the occupational term of a specific person x is γ indicates that x is actually a member of the occupational group G, is a quality of the set of γ. This quality is the distribution of the persons belonging to Γ to specifically organized forms of employment. Calling this distribution of probabilities for belonging to or membership within G μ, we formalize this statement as:

$$x = \gamma \Rightarrow p_G^x = {}^{\mu}\Gamma \quad (4)$$

This distribution of the probability of membership is a quality of Γ, the set of all persons, having an occupation γ; it is causally derived, however, not from other characteristics of these *persons*, but from characteristics of the *environment* – the village, the town – in which they live.

This should be emphasized, as it is of central importance for our following considerations: The probability with which a person with a specific occupational term belongs to a specific occupational category depends on the proportion of members of various socially relevant occupational groups which have been indicated with this term. This relationship between a number of occupational groups indicated by the usage of the occupational term we call a property or quality of the set of all persons for whom this occupational term has been used. For the time being we assume that the "environment" E constitutes a necessary and sufficient causal explanation for this quality: more generally we assume that the *observable* quality is a causal function of the environment. In our formalism:

$$x = \gamma \Rightarrow p_{GG'}^x = {}^{\mu}\Gamma = f(E) \quad (5)$$

Let us use these assumptions to modify our formalism (1), expressing the relationship between one individual pair of fathers and sons to experience a specific type of mobility and the probability for all such pairs, where the sons have the same occupational term, to experience that type of mobility. This modification will relieve us of the necessity to assume that there is an unambiguous mapping between occupational terms and occupational groups. By the basic rules of probability[9] calculus we get the formalism:

$$x = \gamma \Rightarrow p_{GG'}^x = p_{GG'} \cdot p_G^x \cdot p_{G'}^{x'} \qquad (6)$$

From a historical point of view, however, this expression is not without problems: We can assume that the probabilities with which the occupational terms used for father and son indicate their respective status groups are independent from the underlying social process which led to the specific type of mobility they experi-

[9] This paper was to be kept as close as possible to the original notions of 1978, because they are derived from a discussion that should be easily recognizable to anyone familiar with the discussions about occupational coding. Today we would probably discuss how far the various types of 'probability' which are freely combined here do actually allow expressing their dependencies by multiplication. We probably should rather assume that some of them are 'possibilistic' in Zadeh's sense, which has, among others, the effect that they could not be summarized in that way by multiplication. On that concept see: L. A. Zadeh 'Fuzzy Sets as a Basis for a Theory of Possibility', *Fuzzy Sets and Systems* 1 (1978), 3-28.

enced. The assumption that the probabilities with which two persons belonging to various occupational groups are labelled by specific occupational terms should be independent of the co-occurrence of these persons within one social context, is considerably less plausible. Formally we can simply consider this fact by substituting a conditional probability in formalism (6) leading to:

$$x = \gamma \Rightarrow p^{x}_{GG'} = p_{G}G' \cdot p^{x}_{G} \cdot p\left(^{x'}_{G'} \middle|\, ^{x}_{G}\right) \qquad (7)$$

While this modification is formally simple, it makes us almost despair of the sense of our attempts: in plain language we are simply saying that to estimate the probability with which a mobility relation between two occupational groups occurs, we have to estimate the probability with which this very same mobility process influences the mechanism by which a member of a specific occupational group is labelled with a specific occupational term.

Just in case our substantial historical problem has in the meantime been obscured by the formalism used: It is highly probable, that the chance of, for example, occupational stability is different between blacksmiths employed in industrially organized establishments and those working in artisans' shops. Historically, furthermore, it is highly plausible to assume that the civil servant or parish priest producing the lists has been influenced in his choice of vocabulary: either disproportionally frequently employing the same terminology for father and son, though the son experienced a shift to the industrial sector, or being disproportionally more careful with his vocabulary in cases such shifts occurred. In other words: to estimate the chances for occupational stability we have, if we want to consider the degree with which the reality was distorted by the usage of homonyms in our sources, to estimate occupational stability ... Not the last of our problems, unfortunately.

As everybody who is working with material of the type discussed knows, the language of such lists changes, particularly if the regular bookkeeping is taken over by a new civil servant or parish priest: that is, the changes occur discontinuously. This "civil servant or parish priest" whom we encountered as a distorting factor already in the last paragraph has therefore to be considered in our formalisms. Most easily this is done in a very general way: stating that the mapping between occupational term and true occupational categories is not simply a function of the historical environment E, but also of the individual, or more abstractly of the source producing process S. Which modifies formalism (5) to:

$$x = \gamma \Rightarrow p^{x}_{GG'} =^{\mu} \Gamma = f(E, S) \qquad (8)$$

But this formalism gives up the differentiation between observable notation and actual cause, which led us originally to the introduction of $f(E)$ in the right part of formalism (5) after the implication sign. That is, because the properties of the source we are struggling with, and the language of the author responsible for it, are phenomena which we have to consider in the interpretation of the available material; they are however *not* related to the causal chain which created the phenomenon we want to explain – that is, occupational stability.

To solve this deficiency we can abstract our concept one step further, by continuing to generalize our formalism to the right of the implication sign: we state, that the probability with which a person of whom we know the occupational term belongs to a specific occupational group is in a general way proportional to the quali-

ties of two sets. On the one hand this probability depends on the distribution of the occupational groups that may be indicated by the occupational term γ, on the other hand from the general linguistic characteristics[10] σ of the set Ω of all relevant occupational terms within the source available. These two qualities of the observable sets of information are then related in turn in a general way to a causal mechanism in the social environment to be researched. That is:

$$x = \gamma \Rightarrow p_{GG'}^{x} : {}^{\mu}\Gamma,^{\sigma}\Omega : f(U) \quad (9)$$

Bringing our generalization of the problem one step further, we finally reach a very abstract presentation, which has in the opinion of the author to replace the assumptions presented in the formalisms (1) and (2), if we want to use material of the form discussed. We say finally, that the fact that a person has a job which is indicated by a specific occupational term implies that he belongs with a measurable probability to a specific occupational group. This probability is related in an *a priori* unknown way to the qualities of the set of observable entries within the source material, which qualities in turn are related in an *a priori* unknown way to the environment which is mapped by it. Or:

$$x = \gamma \Rightarrow p_{GG'}^{x} : {}^{*}\Omega : f(U) \qquad (10)$$

Here we have performed a step which needs to be explained:

In formalism (10) we summarize a quality of the set of all persons sharing the occupational term γ as part of the properties of all persons which are addressed by any term in the source material. Basically this seems to be acceptable – according to our definition of the quality of a set, it was a property which could not be observed by looking at a single member of that set only. A subset of a set consisting by definition of elements contained in the superset, it is a quite convincing corollary of that basic definition to call the fact that a set contains a subset with a certain quality, a quality of the set containing the subset.

Precisely this step, in the opinion of the author, makes this concept of qualities of a set useful when considering source material of the type discussed. That will be described with a short example.

Analyzing material where we ourselves supervise the collection of information, we can assume that information relevant to our topic of research will appear in the form of unambiguous questions in our questionnaires, which in due time are turned into unambiguous variables of our data sets. Typical examples are questions and variables, which inform us about the occupation of a person and the position of him or her within that occupational field. When we use process produced data, on the other hand, we are concerned with terms which cover more than one of these concepts: the term "blacksmith journeyman", for example, provides information on both the occupational field of the person and his position within it.

At first sight it seems to be extremely easy to derive two variables from this term. Unfortunately we encounter again the problems of individual use of language by the persons creating the sources, which we have already discussed above. In a

[10] On the relationship between semantics and set theory see, for example, Herbert F. Brekle *Semantik,* München 1972, 30 ff.

concrete case it can easily be that between various points in time[11] the following difference can be observed.

Rule #	between	the following holds true
1	t_1 and t_2	Only "blacksmith" occurs, no indication of the individual's position is given.
2	t_2 and t_3	"Blacksmith" is always qualified by "master" or "journeyman", as long as they are working as artisans.
3	t_3 and t_4	"Blacksmith" is sometimes qualified with "master", if the person is self-employed.
4	t_4 and t_5	"Blacksmith" is sometimes qualified with "journeyman", if the person is not self-employed.

In our terminology this system of rules would first be a quality of the source material in general, that is a quality of all the occupational terms Ω occurring within it. Furthermore, it of course would also be a property of our set of all occupational terms Γ – if we modify our initial definition of "γ = blacksmith" to "γ = all phrases containing 'blacksmith' and an optional phrase, explaining about the position of the person in his job."

Beyond that, the individual rules would also be qualities of specific subsets of Γ, the set of all persons designated as γ. If we make our formalism a bit more explicit now, by writing γ_i for the ith person indicated as γ, and with $t_{\gamma i}$ the point in time when the ith person has been assigned that occupational term in our source, we could say that the quality σ (in formalism (9) introduced as linguistic properties) of the set of terms Γ would be given by:

$$^{\sigma}\{\gamma_i | t_1 \leq t_{\gamma i} \leq t_2\} = \text{Rule 1}$$

$$^{\sigma}\{\gamma_i | t_2 \leq t_{\gamma i} \leq t_3\} = \text{Rule 2}$$

$$^{\sigma}\{\gamma_i | t_3 \leq t_{\gamma i} \leq t_4\} = \text{Rule 3}$$

$$^{\sigma}\{\gamma_i | t_4 \leq t_{\gamma i} \leq t_5\} = \text{Rule 4} \qquad (11)$$

Going one logical step further, we discover that our "terms" designating occupations are themselves sets of more basic items of information, which ultimately become terminal symbols in the linguistic sense. These again have qualities which are in a specifiable relationship to the rules we described narratively in the previous discussion.

In our example every member of the set of all occupational terms which signify "blacksmiths", is a set of exactly one element – the occupational term itself – during the applicability of rule 1. During the validity of rule 2, every member of that set is a set of exactly two elements, the term designating the occupational field and the term indicating the position within it. Even if the second element is null, that is,

[11] I am, in general, much indebted to David J. Bartholomew, Stochastische Modelle für soziale Vorgänge, München/Wien 1970. Particularly to the introduction of time dependent mobility models on pages 45-6.

does not exist linguistically, we still are exactly informed about the position of the individual: he is employed in his occupation as a factory worker.[12]

A particularly interesting case exists during the time of applicability of rules 3 and 4: during that time each element of the set of occupational terms indicating blacksmiths is also a set of two elements. If here the second element is null, however, all we know is that this individual did *not* have a specific position. We do not know by that which one he *had*. Including these considerations into our formalism, we get:

$$\begin{aligned} &{}^{\sigma}\{\gamma_i | t_1 \leq t_{\gamma i} \leq t_2\} \Rightarrow \gamma_i = \{\alpha_i\} \\ &{}^{\sigma}\{\gamma_i | t_2 \leq t_{\gamma i} \leq t_5\} \Rightarrow \gamma_i = \{\alpha_i, \pi_i\} \end{aligned} \quad (12)$$

Some explanations are necessary: we replaced the equals sign of formalism (11) more correctly by an implication arrow, as the expression to the right of it represents just part of the rules we defined in the text – the second line within it represents rule 2, 3 and 4, without differentiating between them. The newly introduced symbols α and π represent the linguistic terms for the actual occupational field and for the indication of the position within that field.

Let us review some of the statements developed before in the light of formalism (12)!

First, we can remove a restriction we tacitly made to simplify our argument when we presented formalism (1). There we should, of course, have qualified the presented implication by specifying it as valid only under the assumption that the probability of a mobility transition between the occupations of father and son depends upon the *occupation* of the father only. If we wanted to drop a simplification like that completely, we would have to present a formalism which includes all potentially existing influences if we wanted to say anything about the relationship between two occupational groups. How far this assumption makes sense with conventional social science data remains open – in our case it would simply be absurd. What we should do, however, is to consider all that information in our source which describes the position within the overall social structure, beyond the actual occupational terms *A*.

On the theoretical level we can say that formalism (1) is valid if we replace formalism (2) by one in which we ensure that both the actual occupational term and the linguistic indicator for the position within the occupational field are straightfor-

[12] The author admits that in 1995 one would probably try to get a more elegant conceptualization by applying the concept of "linguistic variables", as formalized by L. A. Zadeh in: L. A. Zadeh 'The concept of a linguistic variable and its Application to Approximate Reasoning' I *Information Science* 8 (1975), 199-249 and II *Information Science* 8 (1975), 301-57. Many of the considerations there deal exactly with the way in which the interpretation of a linguistic term can be influenced by another one. While many of these considerations deal with 'hedges', effectively fuzzy quantifiers like 'many' (E.g.: L. A. Zadeh 'A Fuzzy-Set-Theoretic Interpretation of Linguistic Hedges' *Journal of Cybernetics* 2, 3 (1972), 4-34; G. Lakoff, 'Hedges: a Study in Meaning Criteria and the Logic of Fuzzy Concepts' *Proceedings of the Chicago Linguistic Society* 8 (1972), 183-228), they have occasionally been generalized to the more general concept of linguistic classifiers, which is relatively close to our case: A. K. Nath et al. 'On Some Properties of a Linguistic Classifier' *Fuzzy Sets and Systems* 17 (1985), 297-311.

wardly valid indicators of two types of positions within the overall social structure. That is:

$\alpha \in A, \forall\, \alpha$

$\pi \in P, \forall\, \pi$ (13)

This modification means, however, that we would have to review all our considerations about the most appropriate formalisation of the occupational terms to apply also to the terminology dealing with the position within an occupational category. This is fairly easy: in fact, the rules 1-4 are nothing else, than a proposal, how to estimate the importance of the appearance of a specific term indicating the position within a specific occupational field. This does *not* help us to a straightforward solution of the logical circle which we discovered in formalism (7) and left unsolved for the time being.

The rules we have assumed in the meantime about the treatment of additional occupationally relevant terminology, *did* slightly improve our possibilities for the solution of the problem mentioned there. For the time of applicability of rule 2, the data have removed our problem: the additional information available there solves it, because it simply tells us which occupational group is actually meant by the basic occupational term. If the element π of the phrase designating the occupational position is null, that is, if the phrase contains neither "master" nor "journeyman", we encounter somebody employed in the industry; otherwise we are dealing with an artisan. If we designate these two sets of occupational positions as *I* (Industrial) and *A* (Artisans) respectively, subsets of *G* containing all positions for somebody designated as "blacksmith", we get the trivial case:

$(\alpha_x = \gamma) \wedge (\pi_x = 0) \Rightarrow p^x_{IG'} = p_{IG'}$ (14)

And, replacing *G'* for father's occupation similarly, we get:

$(\alpha_x = \gamma) \wedge (\alpha_{x'} = \gamma) \wedge (\pi_x = 0) \wedge (\pi_{x'} = 0) \Rightarrow p^x_{II'} = p_{II'}$ (15)

This "solution" was unfortunately just made possible because we discovered in part of our data more information than we originally assumed to have. However, for the periods reigned by rules 1, 3 and 4 our original problem has scarcely been tackled yet.

In the opinion of the author, we get closer to a first approximation of a solution if we introduce the following axiomatic assumptions about the nature of information within the described form of social systems.

If the qualities of a set of linguistic items of information about events within a social structure change in discontinuous steps, it cannot be assumed that this change of the surviving information shows a change of the actual social system mapped into that information at the time of the change. If the number of terminal items of information within each element of the subset $\{\gamma_j \mid t_n <= t_{\gamma i} < t_m\}$ *is larger than within each element of the subset* $\{\gamma_j \mid t_k <= t_{\gamma j} < t_l\}$*, it can be assumed that whatsoever cannot be observed in* $t_{\gamma j}$ *by lack of information will by tendency behave very similarly to the same phenomenon at point* $t_{\gamma i}$ *when it can be observed. This the more so, the closer the two points in time are.*

This statement is at first not more than a slightly more formal expression of the common sense assumption that the replacement of a civil servant responsible for keeping lists will usually not be related to a change of the economy or the social

structure of that community: so changes that are observable after the new person takes charge reflect his or her habits, not a change in the "real" relationships. Beyond that the author thinks that this rather axiomatic statement opens up the road for the analysis of "deficient" data of the kind we have discussed here. To do so practically would, however, require a more stringent treatment of two more questions, which go beyond the aims of this paper. Our formalisations do so far not include anything about the possibilities to measure the "closeness" of two points of time in the above assumptions, nor about the relationship between increasing temporal differences and the development of similarities between stages in the development of a social system.

Both problems would be very important for a practical application of these considerations; we can ignore them, however, if we try to draw conclusions from the axiomatic assumption presented.

In our concrete example it simply says that we can use the observable probabilities for the relationship between an occupational term and an occupational group, which we have got from the period during which rule 2 was valid also during the periods which are governed by the other rules; at least in an approximate way. Which is quite similar to what we already did at the beginning, when we entered a distribution of frequencies derived from sources into our formalism (3).

Quite similar, but not identical. Because for our formalism (3) we have drawn upon "additional" information about the social structure we want to investigate; "additional", because it is not contained in the source on which we concentrate. *Now* we just operate with the assumption of a certain degree of continuity. Our axiomatic assumption makes a statement about a quality of the actual social structure, not about the observable information about it. Effectively we could formalize it as the statement that the two following formalisms are *not* contradictory:

$$^{\sigma}\Gamma_i \neq \quad ^{\sigma}\Gamma_j, t_{gi} \rightarrow t_{gj}$$

$$G_i = G_j, t_{Gi} \rightarrow t_{Gj} \qquad (16)$$

Γ_i being the set of all occurrences of *gamma* at t_i, G_i the set of the corresponding occupational positions in reality, also at t_i.

A practical application would of course create additional difficulties. If we want to solve the dilemma which has been described in formalism (7) for the periods during which rules 1, 3 and 4 are valid, we have to assume that the (non-)inheritance of a specific occupation is governed by the same probabilities we could observe during the period in which rule 2 was valid.

A first attempt at operationalization could simply be to insert into our formalism (6) always those probabilities for the indication of a specific occupational group *G* by a specific occupational term γ which we have found during rule 2; to insert them, that is, also for data which describe the periods governed by the other rules. This would give us the necessary mapping from γ to G, which we would need to continue with formalism (1); in other words, to perform a classical mobility study.

That this is legal under our axiomatic assumptions is clear: they say nothing else but that the social system which has historically been mapped into the surviving sources has a tendency to remain stable, even if the process of mapping it into a specific set of sources changes. Nothing but this mechanism entered formalism (5) as $f(E)$. The quality μ of the set of occupational terms γ, that the persons to which

they are assigned belong to certain occupational groups according to known proportions, which we encountered in (5), is part of those qualities of all γ to follow a specific linguistic behaviour – which we have introduced as σ. And this σ is, according to formalism (16) exactly what may change, though the underlying reality remains stable.

Furthermore our axiomatic assumptions allow us to choose the following procedure:

If during the period governed by rule 2, we discover that two linguistically expressed facts are in a specific relationship with each other, we have the possibility to assume that that specific relationship already existed, before the linguistic convention existed and continued to exist after the linguistic conventions could not be observed anymore. More concretely: if during that period sons with occupational information consisting of the elements *{a, b, c}* always co-occur with fathers with *{a, d, e}*, we may rightfully assume that this factual relationship continues to exist when the linguistic rules change[13]. If in such a period we find a son who is described only by *{a}* with a father only described by *{d}*, we may assume that the two would actually more correctly be described as *{a, b, c}* and *{a, d , e}* also in this period, where the information provided is more sparse. (Of course, "always co-occur" can be replaced by "co-occur with an observable probability".)

The approach described here is actually neither very complicated, nor particularly original: it just tries to formalize assumptions which in historical research have been used time more or less tacitly for quite some time. – Even this simple formalization allows us to extend the domain of research based upon statistical reasoning to new classes of information, which could not be processed without these assumptions. To prove this is the purpose of the remainder of this paper.

Because of the central importance of that concept within this paper we would like to clarify further what we mean by a "linguistic property" of a source. The "linguistic properties" of a source specify the total sum of all the rules which govern the way in which "information" – in the sense given above – is used to express real social relationships and structures.

1.2 Applications and Consequences

If we try to apply the preceding considerations, we have to solve two problems. First: we have to try to filter out of a collection of highly fuzzy sets of sets of information variables which are defined clearly and precisely enough for use in formal computations. If this is successful, we have to show that such a way can be used in actual research, i.e. that there are computational techniques which abbreviate the long road from fuzzy terms to statistical results sufficiently to make it viable within normal research projects.

To pinpoint the first of these two problems: speaking strictly statistically, we could say, that we are working with variables below the nominal level of scaling.

[13] Here we intentionally abandon our convention of using Greek characters for linguistic terms and Latin ones for actual social phenomena: the following would be true for relationships between linguistic terms as well as between elements of the 'real' social position for which we do or do not have 'hard' information.

We know that a specific term may contain some information and cannot contain some other information. We can, however, not be sure that one and the same term will always express exactly the same concept when it is encountered in different contexts – for which the occurrence in different periods is just one, though the most obvious, example.

If we postpone the details of implementation, the way to go for a research project is clear after the preceding theoretical description. We have first to isolate that part of our source which contains the "best" mapping of social reality into linguistic terms – that is that part, where the terms contain the most information. By analyzing the relationships between these items of information, we can derive probabilities for the general mapping from underlying reality into linguistic terms (and an attempt to invert it).

We can, for example, assume that the following data exist, where "I" shall represent persons who have a specified occupation and a specified status within it. "II" represents persons, who have the same occupation, but another status within it, and "III" persons who have another occupation.

Table 1

Fathers	Sons		
	I	II	III
I	a	b	c
II	d	e	f
III	g	h	i

In a later stage the information on the position within the occupation is missing completely – we are left with the following items:

Table 2

Fathers	Sons	
	I, II	III
I, II	x	y
III	z	w

Our statement that, despite these changes in the linguistic behaviour, we still have a similar causal structure underlying that table, implies that we assume that behind these observed values the following ones can be assumed to have existed:

Table 3

Fathers	Sons		
	I	II	III
I	$\frac{xa}{a+b+c+d}$	$\frac{xb}{a+b+c+d}$	$\frac{yc}{c+f}$
II	$\frac{xd}{a+b+c+d}$	$\frac{xc}{a+b+c+d}$	$\frac{yf}{c+f}$
II	$\frac{zg}{g+h}$	$\frac{zh}{g+h}$	w

In this form it looks like a poor result to get from a ten-page formalization. In the opinion of the author it becomes considerably less trivial, however, if we emphasize

that this is not a deterministic continuity, but a tendency towards continuity; a probabilistic relationship, in other words, which we will explore further later. And such a probabilistic relationship does not only govern this specific example, but the process of mapping the actual structures and events into observable information in general.

These two aspects of our considerations do not say anything other than that we should also expect the same continuity in parts of the accessible data for which we have roughly the same amount of information during the validity of both rules of mapping. This means that in our Table 1 for the field *a* the following has to be true:

$$a_{t_2} \sim \frac{\left(a_{t_2}+b_{t_2}+c_{t_2}+d_{t_2}\right)a_{t_1}}{a_{t_1}+b_{t_1}+c_{t_1}+d_{t_1}} \quad (17)$$

Whether this is the case can easily be checked; the degree to which the actual values at t_2 deviate from the ones which we have obtained with this kind of formula, is a measurement for the degree of stability in the underlying social structure. And, assuming that the degree of stability is the same for all structures and processes we are interested in, we can now use this value to weight the values in our Table 3.

One would presumably try to compute this degree of stability for as many relationships which are expressed by constant amounts of information as possible: by which method these almost certainly different degrees of deviation should be summed up into a single value is one of the two problems which we would think of immediately. The second one is a bit more complex: in our axiomatic assumption we stated that the actual positions described by terms remain unchanged by tendency. How precisely such "positions" can be defined, we have not discussed at all.

A practical example: we have assumed that it is legal to draw conclusions about the position of blacksmiths within their field of occupation from data which are unevenly precise over time. If in $t_1 \rightarrow t_2$ we have information about that position, we assumed that in $t_2 \rightarrow t_3$ the distribution of such positions would roughly be the same. So much for blacksmiths. But how about tailors? Our last formal proposals imply that the degree of deviation from the previous distributions which we might observe for blacksmiths would also hold true for tailors.

Going back to the conceptual level: how far can a common "position" between these two occupational groups be construed, which would allow us to estimate linguistic changes related to both of them?

The first of the two questions just raised – the question how observable rates of deviation from the original distributions can be reasonably be summarized – is as difficult as it is, because it more or less asks of us to find a solution for how to summarize data on the nominal level (on which occupational information is solidly based). The plain answer is simple: not at all.

This seemingly completely negative answer forms a starting point, however, for us to put our still highly theoretical considerations to practical work. All our difficulties go back to the fact that we tried to derive abstract principles from concrete linguistic examples, where it was not really clear how they related to each other. What we have been successful at, as in the case of the blacksmith example, has been not more than the derivation of nominal information from, as we called it, subnominal data.

From the historical point of view, that is not quite as little as it may look like to a social scientist, who is used to operating with data sets which claim to enable him or her to draw conclusions about societies as a whole. From that background the possibility to create very specific statements about singular and highly local developments in periods where it would not have been possible before, seems to be unimportant indeed.

Precisely this, however, seems to be what our considerations are leading us towards to. That is, we have consciously to avoid *general* quantitative statements about "mobility" if we have data material of the kind described. Another, admittedly less elegant, way remains open, however: we can explore specific "mobility transitions" and test, with the help of the strategies we discussed last, how far the results we get for such very narrowly defined transitions between well-defined positions can be generalized to other groups within the society in which they are embedded.

At that stage, however, the application of statistical methods is absolutely necessary. Let us assume that we finish our considerations about occupational inheritance with the occupational group of blacksmiths for a longer period within a specified area. In such a well-defined environment, we obviously have the possibility to explore a whole range of formal linguistic properties, their co-occurrences and significance. With the proposed procedure, we than *have* the possibility to derive a description of occupational stability with the groups of blacksmiths over a pretty long period.

Not only for them, but similarly also for other such narrowly defined groups. If that is done, nobody prevents us from comparing the individual reconstructions. They certainly will not show exactly the same trends – in all probability, we will, however, discover similarities in the developments of individual groups.

"Similarities" have, of course, also been derived earlier: indeed a large part of classical social history simply attempts to derive plausible generalizations from individual indicators and present them as narrative. The consistency of these individual generalizations can scarcely be without doubt, however, and there is usually no way to discuss it on a truly inter-subjective level.

Our proposal, on the other hand, would allow us to quantify such generalizations and, much more importantly, to formalize tests for their consistency. The approach we presented here provides us with measurable parameters, which map properties of the linguistic material; proportions which allow us to predict how linguistic phenomena must have been generated, if we have correct assumptions about the underlying social processes. Such proportions will of course depend very much on which aspect of a very complex overall process we discuss – but whether such differences are arbitrary or within an expected range would be an excellent field for the application of statistical testing.

Not necessarily for the testing of its more conservative type; why we should assume, that the transitional probabilities of a series of occupational groups, which we might interpret as a sample out of the universe of all existing occupational groups, should be distributed normally, is, for example, completely unclear. On the other hand, data of this kind would form an excellent test for parameter free methods.

3. Two Angry Brothers: Quantification and Source Oriented Computing, AD 1993, or: Why are we not a hard science?

The original internal paper, from which we have given an abbreviated and slightly revised version above, goes on for a couple of more pages, describing the new activities of the author than just recruited by the *Max-Planck-Institut für Geschichte* as, among other things, an attempt to create a software environment, to make quantitative research with data of the described type possible within history, by providing tools for iterative coding procedures implied by the above. A software environment, which eventually became known as CLIO and later Κλειω. If one knows its origins there are considerable traces of these early ideas in the system: it is probably the only existing data base system, where even in the fundamental data model[14] all numerical comparison operators exist in fuzzy versions (implementing things like an "roughly equal or larger" operator, which accepts numbers like "approximately 50" or "less than 40, but possibly a little bit larger" as operands). Some of the initial considerations have even spawned an independent research project, which formulated the notion of a *context sensitive data base*[15] and turned that into a working experimental implementation[16].

Still, very few users of Κλειω would recognize the reworked paper from which we have started as a design scheme for the system they are working with. And more recently the author tends to describe the difference between historical and contemporary disciplines, or that between the Humanities and the hard sciences, in a much more pointed way, as follows in the form of another lengthy quotation from an address delivered to an audience almost exclusively made up of hard science researchers.[17]

We probably all agree, that the discovery of the equations of Gay-Lussac at the start of the last century are one of the more important milestones on the way to the hard sciences as we know them. If we accept the famous $pV = p_0 \, V_0(1+\alpha t)$ as a typical example for a hard science statement, we could be tempted to ask why the historical research community has never arrived at anything of even remotely similar precision in their field – for example a clear law governing the relationship of nutritional situation and frequency of births within a given historical society. On the

[14] Not to claim that fuzzy extensions of other data models and data base systems do not exist for more conventional ones: cf. P. Bose et al. 'Fuzzy Querying with SQL: Extensions and Implementation Aspects' *Fuzzy Sets and Systems* 28 (1988), 333-49 or already Billy P. Buckles and Frederick E. Petry 'A Fuzzy Representation of Data for Relational Databases' *Fuzzy Sets and Systems* 7 (1982), 213-26.

[15] Cf. Manfred Thaller 'Databases and Expert Systems as Complementary Tools for Historical research' *Tijdschrift voor Geschiedenis* 103 (1990), 233-47

[16] Wolfgang Levermann: Kontextsensitive Datenverwaltung, St. Katharinen: Scripta Mercature 1991 (= Halbgraue Reihe zur Historischen Fachinformatik) B8.

[17] M. Thaller 'Die Herausforderung großer Korpora unstrukturierter Texte', *Forschung und wissenschaftliches Rechnen* (= Max-Planck-Gesellschaft, Berichte und Mitteilungen 1/94), 32-44, here: 32-3.

level we have just quoted, that question almost answers itself: at least the "nutritional situation" of a society is an abstraction of several more orders of magnitude than the pressure within a gas. At a lower level, say the relationship between the financial situation of a family, its relative social rank and the number of children it spawned, one could assume we are much closer to observable units. And, one could argue, the health and personal decisions of individuals could just as well be explained as arbitrary as the "decisions" of individual atoms in a gas – and therefore possibly overcome by statistical methods.

Most historians would simply point out that even seemingly simple concepts in historical communities are actually quite complex. We would like, however, to show a much more fundamental difference between the two groups of disciplines. If we accept Gay-Lussac's law as the model of a hard science statement, it can obviously be generalized to $c_1, c_2, \ldots c_i \Rightarrow E$ (a series of observable causes has an observable effect). Or, as the most skeletonized form a hard science statement can take, *observable causes have observable effects:*

$$C \Rightarrow E$$

The real problem is hidden in the innocent word "observable". What a thermometer tells us about a gas is indisputable: whether the tax, which is assigned to a family in a list of taxation indicates its financial status, or the social rank it was assigned by an older type of bureaucracy, whether that tax even has ever be paid, is completely open. The equivalent of a statement of the historical sciences, on the highly abstract level chosen above, would therefore have to be:

$$\begin{array}{ccc} A & \Rightarrow & B \\ \downarrow & & \downarrow \\ \alpha & & \beta \end{array}$$

Two historical "facts", α and β, have survived. Assuming, that the hypotheses are correct, how two events in reality, A and B, have been mapped (↓) into the "facts" α and β, there has been a causal relationship between A and B.

Which means: Even on the most extreme level of abstraction, the most simple statement in the historical disciplines is a statement about the consistency of three hypotheses, two related to the tradition of information, one to the relationship between the actual events.

What would have become of the hard sciences if Gay-Lussac had never been sure whether the instruments measured the temperature or rather more the magnetic field within the gas?

So far a presentation of our case to a group of hard scientists. In many ways this is a very consistent summary of the more detailed treatment of the question in the previous pages of this paper: one could call the proposition just presented a generalization of formalism (7) pointing towards a logical circle in the earlier presentation, out of which we proposed there to break with the help of an iterative procedure.

The great difference lies in the fact that in 1978 the purpose was to lay the groundwork for a quantitative solution; in 1993 we have shown the problem just to explain, why quantitative methods were not quite that central to historical research. In that move away from statistics this author has not been alone: as is easily shown by, for example, the history of nominative record linkage. In the days of the Phila-

delphia History project, one of the more central concerns has been to develop a complex system of probabilistic weights to decide between equi-probable links[18]; and in those years almost all research in that area was involved in similar considerations. In the meantime, nominative record linkage deals predominantly with rules-of-thumb about how to get the information contained within two lists together – though the more prominent of its practitioners will still present a position towards weighting systems, which, however, is scarcely relevant for the actual practice of the craft.

Where does this development come from?

The author obviously knows best why, in the area where *he* has been working, the once ever-present feeling that one wanted to prepare source material for a final, crowning, statistical analysis has diminished. He has the strong suspicion, however, that that has been a quite general experience. Without any claim for completeness, there are at least the following explanations for the fact that source oriented computing has appeared to separate itself more and more from quantitative methods.

Problem 1. In the seventies and early eighties, with all the excitement about the increasing computer power on one's desk, some of the steps involved in implementing the proposals in the first part of this paper have simply needed more resources than were easily available. The reason why the prototype of a context sensitive version of Κλειω never made it into the production version was quite simply that under the prevalent feeling "everything has to work on a PC", these functions were considered not yet to be applicable to the typical user. And if they were hard to realize: was it even necessary, when so many people were enthusiastic at being able to use a computer without learning about Pearson's existence? Should you on the other hand press them towards statistics, when you saw valid reasons why that might not be all that applicable to their material in the first place?

Problem 2. In cases where source material is genuinely quantitative, statistical methods used to draw conclusions from that material can be meaningfully applied to relatively small datasets, to datasets, that is, which can be collected easily by individual researchers. However, this is not the case if we try to look statistically at problems which are not documented by genuinely quantitative material. In the first versions of the record linkage tools of CLIO, the Philadelphia History Project weighting system was carefully implemented. After two years or so it has been removed – as the experience with a number of projects showed that it would actually simplify the tasks only if you had more than at least 25.000 individuals to identify. The reason why the author of this paper could produce his cluster analysis of the influence of the theme of medieval images upon the colouring scheme used by the artists[19] was that a huge data base with descriptions of such images existed – which had been created for much more mundane things. But for many other of the theoretically interesting things, you could either use datasets so small that other historians did not find the result interesting, or you would have to spend ages in data collection.

---

[18] Theodore Hershberg et al. 'Record Linkage' *Historical Methods Newsletter* 9, 2-3 (1976), 137-63.

[19] Thaller 'Zur Formalisierbarkeit hermeneutischen Verstehens in der Historie', as above.

4. Two Aging Brothers: Quantification and Source Oriented Computing, AD 2001, or: Where are we Headed?

From the point of view of this author, two scenarios for the relationship of two fields of work which grew up in such close relationship are possible:

Scenario 1: Quantitative studies become the accepted tool for the analysis of sources, where the information is arguably precise and consistent. The rest of the discipline ignores these methods as clearly inappropriate for their purposes.

Scenario 2: "Quantitative" methods concentrate more on the question, under which conditions they can be applied to source material where they have not be applied so far. Source oriented computing helps to implement this.

This author considers the second of the two to be much preferable.

The situation for such a development would be considerably better today than it was twenty or even ten years ago. The computer technology of the late nineties finally seems to keep the promises that the enthusiasm of the early years of the "micro-revolution" made. So the actual lack of computing power, which has been given as one reason why some of the methodically more challenging concepts of source oriented computing have not made their way into actual applications, is finally in the process of being resolved. At the same time the size of historical data sets is virtually exploding: one of the really interesting questions of the next few years will, for example, be whether quantitative content analysis will leave the stagnation it has unmistakably entered some years ago. During the last decades these studies were always based upon corpora of texts, where only in very isolated cases quantitative methods could discover details which were completely beyond the ability of a researcher to discover unaided, as the "corpus" of texts quite frequently was restricted to a few hundred pages. Today the rapidly expanding collections like the *Patrologia* provide us with a plethora of textual material. Will source oriented processing rise to the challenge to make these materials accessible with something more sophisticated than fulltext retrieval systems, the relevance of which source oriented computing discussed in the early eighties? Will quantitative methodology rise to the challenge?

1) We need the courage to embrace statistical methodology which is appropriate for history, even if it is not available in ready-made form. The contributions on fuzzy methods in this volume I consider most promising in that respect.
2) We have to move away from the concept that somewhere out there exists a methodological canon, which we simply have to learn and to apply to our material. We have to study ourselves which formal properties historical sources have.
3) We have to implement tools which help us to transform historical sources so that probabilistic answers can be arrived at. Such software tools would have to implement the kind of reasoning with which this paper started.
4) If some of our colleagues are sceptical about statistics because they see no "variables" in the bewildering material they are used to handle, it will not be sufficient, to show them that in other material such variables exist. – Particularly not, if an ever increasing arsenal of data base and similar tools make it deceptively easy to handle that same kind of material on a (seemingly) informal level.